On the Stability of Gauss-Jordan Elimination with Pivoting

G. Peters and J.H. Wilkinson
National Physical Laboratory
Teddington, Middlesex, England

The stability of the Gauss-Jordan algorithm with partial pivoting for the solution of general systems of linear equations is commonly regarded as suspect. It is shown that in many respects suspicions are unfounded, and in general the absolute error in the solution is strictly comparable with that corresponding to Gaussian elimination with partial pivoting plus back substitution. However, when A is ill conditioned, the residual corresponding to the Gauss-Jordan solution will often be much greater than that corresponding to the Gaussian elimination solution.

Key Words and Phrases: Gauss-Jordan algorithm, Gaussian elimination, back-substitution, backward error analysis, bounds for error in solution, bound for residual
CR Categories: 5.11, 5.14

1. Introduction

The essential numerical stability of Gaussian elimination with partial pivoting is commonly demonstrated by the technique of backward error analysis [1, 2, 3, 4]. Such an analysis shows that, when an $n \times n$ system is solved on a computer working in floating-point arithmetic in base β with a t-digit mantissa, the computed solution x_c is the exact solution of some "neighboring" system

$$(A+E)x_c = b. \tag{1.1}$$

The term "neighboring" is used in a rather loose sense. In fact, a bound is found for an appropriate E, which is of the form

$$\| E \|/\| A \| \leq f(n)g\beta^{-t}, \tag{1.2}$$

where l_2 or l_∞ norms are commonly used, $f(n)$ is a modest function of n, and g is the "growth" factor. The latter is defined to be the ratio of the coefficient of maximum modulus occurring during the course of the elimination to $\max | a_{ij} |$. The role played by pivoting is in limiting the probable growth. Although, even with partial pivoting g *can* attain the value 2^{n-1}, in practice it is commonly of the order of unity. For systems of order greater than 10, the statistical distribution of the rounding errors usually ensures that $f(n) \not> n$.

Let us analyze for the moment the consequences of such a result. Suppose for example

$$\| E \|_\infty/\| A \|_\infty \leq n\beta^{-t}. \tag{1.3}$$

If we write $\kappa = \| A \|_\infty \| A^{-1} \|_\infty$, then provided $n\beta^{-t}\kappa < 0.1$ (say) the relations (1.1) and (1.3) ensure that

$$\| x_c - x \|_\infty / \| x \|_\infty \leq n\beta^{-t}\kappa / (1 - n\beta^{-t}\kappa) \leq (\tfrac{10}{9})n\beta^{-t}\kappa. \tag{1.4}$$

The accuracy of the computed solution is therefore directly dependent on κ, the *condition number* of A. The relations (1.4) imply that

$$[1 - (\tfrac{10}{9})n\beta^{-t}\kappa] \| x \|_\infty \leq \| x_c \|_\infty \leq [1 + (\tfrac{10}{9})n\beta^{-t}\kappa \| x \|_\infty \tag{1.5}$$

or, from our assumption that $n\beta^{-t}\kappa \leq 0.1$, certainly that

$$(\tfrac{8}{9}) \| x \|_\infty \leq \| x_c \|_\infty \leq (\tfrac{10}{9}) \| x \|_\infty. \tag{1.6}$$

Hence $\| x_c \|_\infty$ is certainly of the same order of magnitude as $\| x \|_\infty$, and when $n\beta^{-t}\kappa$ is much smaller than unity, the two norms will be almost equal. On the other hand, the residual vector r defined by $b - Ax_c$ satisfies the relations

$$\| r \|_\infty = \| b - Ax_c \|_\infty = \| Ex_c \|_\infty \leq \| E \|_\infty \| x_c \|_\infty \leq n\beta^{-t} \| A \|_\infty \| x_c \|_\infty. \tag{1.7}$$

In other words, we have a bound for $\| r \|_\infty$ which depends only on the *size* of the computed solution and not upon the condition number of A and therefore not upon the *accuracy* of x_c. The errors in x_c are correlated in a way which ensures that r is normally much smaller than might be expected when κ is large. Indeed if we take a random vector x_c satisfying condition (1.4), then for such an x_c one can guarantee only that

$$\| r \|_\infty = \| b - Ax_c \|_\infty \leq \tfrac{10}{9} \| A \|_\infty n\beta^{-t}\kappa \| x \|_\infty. \tag{1.8}$$

In general approximate solutions of the same accuracy as that given by Gaussian elimination give residuals which are larger by a factor κ. That the computed solution gives such a small residual may be very important in practice. We may well be more interested in the proximity of Ax_c to b than in the absolute accuracy of x_c. To emphasize the extraordinary nature of the correlation, we remark that the residual given

Authors' address: Department of Trade and Industry, National Physics Laboratory, Division of Numerical Analysis and Computing, Teddington, Middlesex, England.

by the computed solution is of the same order of magnitude as that corresponding to *the correctly rounded solution.*

When a backward error analysis of Gauss-Jordan elimination is attempted, it is found that one cannot demonstrate that the computed solution is an exact solution of some "neighboring" system with any reasonable interpretation of the word "neighboring." The failure stems from the fact that, with Gauss-Jordan, pivoting does not give satisfactory control of "growth." Indeed it really is no longer true in general that the computed x_c is the solution of a neighboring system. For this reason Gauss-Jordan is commonly regarded with suspicion by numerical analysts. It is the purpose of this paper to demonstrate that this suspicion is only partly justified.

It should be emphasized that in certain cases such as when A is positive definite or diagonally dominant it is well known that Gauss-Jordan is stable.

2. Description of the Gauss-Jordan Algorithm

Since the Gauss-Jordan algorithm with pivoting is well known, we shall describe it only briefly. We denote the original system by

$$A^{(1)}x = b^{(1)}. \tag{2.1}$$

There are n major steps. At the beginning of the rth step the original system has been replaced by an equivalent system

$$A^{(r)}x = b^{(r)} \tag{2.2}$$

in which $a_{ij}^{(r)} = 0$ $(j=1, 2, \ldots, r-1;\ i \neq j)$. This means that $A^{(r)}$ is diagonal as far as its first $r-1$ columns are concerned. The rth major step proceeds as follows.

(i) Let $\max_{i \geq r} |a_{ir}^{(r)}| = |a_{r',r}^{(r)}|$. (In the case of ambiguity, r' is taken to be the smallest such index.)
(ii) Interchange equations r and r'.
(iii) For each value of $i \neq r$, compute $m_{ir} = a_{ir}^{(r)}/a_{rr}^{(r)}$ and subtract m_{ir} times equation r from equation i.

The final system $A^{(n+1)}x = b^{(n+1)}$ is clearly such that $A^{(n+1)}$ is diagonal.

From the choice of r' it is clear that $|m_{ir}| \leq 1$, $(i > r)$, but for $i < r$ there is no such bound on m_{ir}. This means that although the growth of elements below the diagonal elements is limited as in Gaussian elimination with column pivoting (indeed it *is* Gaussian elimination as far as these elements are concerned), growth of elements above the diagonal may be arbitrarily large. This precludes the possibility of a satisfactory backward error analysis analogous to that for Gaussian elimination.

Fig. 1.

$$U = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ & \epsilon_2 & 1 & 1 & 1 \\ & & \epsilon_3 & 1 & 1 \\ & & & \epsilon_4 & 1 \\ & & & & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & \epsilon_2^{-1} & \epsilon_2^{-1} & \epsilon_2^{-1} \\ & \epsilon_2 & 1 & 1 & 1 \\ & & \epsilon_3 & 1 & 1 \\ & & & \epsilon_4 & 1 \\ & & & & 1 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 1 & 0 & 0 & (\epsilon_2\epsilon_3)^{-1} & (\epsilon_2\epsilon_3)^{-1} \\ & \epsilon_2 & 0 & \epsilon_3^{-1} & \epsilon_3^{-1} \\ & & \epsilon_3 & 1 & 1 \\ & & & \epsilon_4 & 1 \\ & & & & 1 \end{bmatrix}$$

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 & (\epsilon_2\epsilon_3\epsilon_4)^{-1} \\ & \epsilon_2 & 0 & 0 & (\epsilon_3\epsilon_4)^{-1} \\ & & \epsilon_3 & 0 & \epsilon_4^{-1} \\ & & & \epsilon_4 & 1 \\ & & & & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ & \epsilon_2 & 0 & 0 & 0 \\ & & \epsilon_3 & 0 & 0 \\ & & & \epsilon_4 & 0 \\ & & & & 1 \end{bmatrix}$$

3. Standard Backward Error Analysis of Gauss-Jordan

Our remarks show that the part of Gauss-Jordan which is suspect is the production of zeros above the diagonal. It is convenient for our purposes to think of Gauss-Jordan with pivoting as taking place in two distinct stages: (i) the reduction to upper triangular form by the standard Gaussian elimination algorithm with partial pivoting and (ii) the further reduction of the triangular system to a diagonal system by an elimination process in which pivoting is precluded. The reader may easily convince himself that when the computing is done in this order the rounding errors are the same as in the classical procedure. The essential difference between solution by Gaussian elimination and by Gauss-Jordan is that in the former the resulting upper triangular system is solved by back-substitution and in the latter, by a further elimination to diagonal form. We may therefore concentrate on the numerical stability of the solution of an upper triangular system $Ux = c$ by elimination.

An examination of this process with an eye to performing a backward error analysis soon reveals the difficulty. This may be exposed by means of a simple example. In Figure 1 we show the reduction of a system of order 5 to diagonal form. We give only the orders of magnitude of the computed quantities. In the original triangular matrix it is assumed that all elements are of the order of unity except for the diagonal elements u_{22}, u_{33}, u_{44}, which are assumed to be small. We write $u_{ii} = \epsilon_i$ $(i=2,3,4)$.

It will be observed that, except in rare cases when cancellation occurs, considerable growth takes place and elements are derived which are proportional to products of the reciprocals of the ϵ_i. Now in a backward error analysis, the equivalent perturbations resulting from any stage of reduction are directly proportional to the size of the elements which arise in the reduced matrix. Accordingly a backward error analysis shows that the final diagonal set of equations is that which

would have arisen from exact computation with $U + E$ where a bound for $| E |$ is obtained of the form

$$\beta^{-t}\begin{bmatrix} 1 & 1 & \epsilon_2^{-1} & (\epsilon_2\epsilon_3)^{-1} & (\epsilon_2\epsilon_3\epsilon_4)^{-1} \\ & 1 & 1 & \epsilon_3^{-1} & (\epsilon_3\epsilon_4)^{-1} \\ & & 1 & 1 & \epsilon_4^{-1} \\ & & & 1 & 1 \\ & & & & 1 \end{bmatrix}. \tag{3.1}$$

If it *is* nevertheless true that the computed solution is as accurate as can be expected, having regard to the condition U, we cannot expect to establish this via the version of backward error analysis we have just sketched. This situation is in striking contrast to that holding for back-substitution in a triangular system. There it is easy to show that one always obtains an exact solution of some system with matrix $U + E$ where certainly $| e_{ij} | \leq n\beta^{-t} | u_{ij} |$ and hence the small u_{ii} do not adversely affect the matrix E. However, the fact that E is now disappointingly large does not necessarily mean that the solution is *bound* to be correspondingly poor. There will be many sets of equations *entirely* different from $Ux = c$ which have *exactly* the same solution!

We observe that the large perturbations in U are in positions which are specially related to the positions of the ϵ_i. Is it possible that the large perturbations occur in just those positions where they have least effect? We now show that this is indeed true for the example we have just considered. Observe first that the condition of U is at least of order $(\epsilon_2\epsilon_3\epsilon_4)^{-1}$ so that even perturbations of order β^{-t} only are capable of producing relative perturbations in a solution of order $(\epsilon_2\epsilon_3\epsilon_4)\beta^{-t}$. Since we have perturbations of order $(\epsilon_2\epsilon_3\epsilon_4)^{-1}\beta^{-t}$, there appears to be a danger that we shall get relative perturbations of order $(\epsilon_2\epsilon_3\epsilon_4)^{-2}\beta^{-t}$ in the solution. This fear proves to be unfounded. A first order argument will suffice to establish this. We have

$$(U+E)^{-1}c \doteq U^{-1}c - U^{-1}EU^{-1}c = x - (U^{-1}E)x. \tag{3.2}$$

We are therefore interested in $U^{-1}E$, i.e. in the solution F of $UF = E$. Let us consider the last column of F. We have

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ & \epsilon_2 & 1 & 1 & 1 \\ & & \epsilon_3 & 1 & 1 \\ & & & \epsilon_4 & 1 \\ & & & & 1 \end{bmatrix}\begin{bmatrix} f_{15} \\ f_{25} \\ f_{35} \\ f_{45} \\ f_{55} \end{bmatrix} = \beta^{-t}\begin{bmatrix} (\epsilon_2\epsilon_3\epsilon_4)^{-1} \\ (\epsilon_3\epsilon_4)^{-1} \\ \epsilon_4^{-1} \\ 1 \\ 1 \end{bmatrix} \tag{3.3}$$

and it is immediately apparent that in general the orders of magnitude of the f_{15} are expressed by the relation

$$\begin{bmatrix} f_{15} \\ f_{25} \\ f_{35} \\ f_{45} \\ f_{55} \end{bmatrix} \doteq \beta^{-t}\begin{bmatrix} (\epsilon_2\epsilon_3\epsilon_4)^{-1} \\ (\epsilon_2\epsilon_3\epsilon_4)^{-1} \\ (\epsilon_3\epsilon_4)^{-1} \\ \epsilon_4^{-1} \\ 1 \end{bmatrix}; \tag{3.4}$$

no square of any ϵ_i is involved. If the perturbations of order $(\epsilon_2\epsilon_3\epsilon_4)^{-1}$ in E had occurred in positions (5,5) or (5,4), then f_{15} and f_{25} would have been of the order of $(\epsilon_2\epsilon_3\epsilon_4)^{-2}$. There is little point in trying to regularize this approach since the analysis of the next section is much more satisfactory, but we may comment here on one consequence of our result. We have mentioned before that when $(A+E)x = b$ then $r = b - Ax = Ex$ and $\| r \|_\infty \leq \| E \|_\infty \| x \|_\infty$. We have been unable to obtain a small E in our analysis of the solution of a triangular system by elimination. There appears therefore to be a danger that the errors in the computed solution will not be correlated in such a way as to give an impressively small residual.

4. Detailed Error Analysis

The essential numerical stability of Gauss-Jordan may be established by a backward error analysis having a somewhat different objective from that described in the previous section. For the reasons already given, we can restrict ourselves to the consideration of an upper triangular system.

Let us concentrate for the moment on the operations which reduce the first equation $u_{11}x_1 + \cdots + u_{1n}x_n = c_1$ to an equation involving x_1 only. Described in a simplified notation this is achieved by subtracting in succession y_2 times equation 2, y_3 times equation 3, ..., y_n times equation n from the first equation. Forgetting rounding errors for the moment we have

$$\left.\begin{aligned} u_{12} - y_2u_{22} &= 0, \\ u_{13} - y_2u_{23} - y_3u_{33} &= 0, \\ &\vdots \\ u_{1n} - y_2u_{2n} - y_3u_{3n} - \cdots - y_nu_{nn} &= 0, \end{aligned}\right\} \tag{4.1}$$

and the final derived equation is

$$u_{11}x_1 = c_1 - y_2c_2 - y_3c_3 - \cdots - y_nc_n. \tag{4.2}$$

In practice the computed $\bar{y}_i$ and $\bar{x}_i$ are determined by the relations

$$\left.\begin{aligned} \bar{y}_2 &= \mathrm{fl}[u_{12}/u_{22}], \\ \bar{y}_3 &= \mathrm{fl}[(u_{13} - \bar{y}_2u_{23})/u_{33}], \\ &\vdots \\ \bar{y}_n &= \mathrm{fl}[(u_{1n} - \bar{y}_2u_{2n} - \bar{y}_3u_{3n} - \cdots - \bar{y}_{n-1}u_{n-1,n})/u_{n,n}], \\ \bar{x}_1 &= \mathrm{fl}[(c_1 - \bar{y}_2c_2 - \cdots - \bar{y}_nc_n)/u_{11}]. \end{aligned}\right\} \tag{4.3}$$

In other words, $\bar{y}_i$ and $\bar{x}_1$ are derived by solving the triangular system.

$$\begin{aligned} &u_{22}y_2 = u_{12}, \quad u_{23}y_2 + u_{33}y_3 = u_{13}, \\ &\cdots \\ &u_{2n}y_2 + u_{3n}y_3 + \cdots + u_{n,n}y_n = u_{1n}, \\ &c_2y_2 + c_sy_3 + \cdots + c_ny_n + u_{11}x_1 = c_1, \end{aligned} \tag{4.4}$$

by *the forward substitution process*, and we know from the conventional analysis of back-substitution in Gaussian elimination [3] that this process is *very* stable.

Indeed the computed values satisfy exactly equations of the form

$$\begin{aligned}
&u_{22}(1+\epsilon_{22})\bar{y}_2 = u_{12}(1+\epsilon_{12}),\\
&u_{23}(1+\epsilon_{23})\bar{y}_2 + u_{33}(1+\epsilon_{33})\bar{y}_3 = u_{13}(1+\epsilon_{13}),\\
&\quad\vdots\\
&u_{2n}(1+\epsilon_{2n})\bar{y}_2 + u_{3n}(1+\epsilon_{3n})\bar{y}_3 + \cdots +\\
&\qquad + a_{nn}(1+\epsilon_{nn})\bar{y}_n = u_{1n}(1+\epsilon_{1n}),\\
&c_2(1+\epsilon_2)\bar{y}_2 + c_3(1+\epsilon_3)\bar{y}_3 + \cdots + c_{nn}(1+\epsilon_n)\bar{y}_n\\
&\qquad + u_{11}(1+\epsilon_{11})\bar{x}_1 = c_1(1+\epsilon_1). \qquad (4.5)
\end{aligned}$$

We are not interested in the most precise bounds for the ϵ_{ij} and ϵ_i. It will suffice for our purposes to observe that certainly

$$\begin{aligned}
&(1-\epsilon)^n \le 1 + \epsilon_{ij} \le (1+\epsilon)^n,\\
&(1-\epsilon)^n \le \epsilon_i \le (1+\epsilon)^n,
\end{aligned} \qquad (4.6)$$

though most of the ϵ_{ij} and ϵ_i will satisfy stricter bounds; here ϵ is a bound for the relative error made in an arithmetic operation. See, e.g. [3]. (On a typical computer employing rounding, $\epsilon = \frac{1}{2}\beta^{1-t}$). This means that $\bar{x}_1$ is precisely the first component of the exact solution $x^{(1)}$ of the "neighboring system."

$$(U+\delta U^{(1)})x^{(1)} = c + \delta c^{(1)} \qquad (4.7)$$

where certainly

$$|\,\delta U^{(1)}\,| \le n\epsilon\,|\,U\,|,\ |\,\delta c^{(1)}\,| \le n\epsilon\,|\,c\,|. \qquad (4.8)$$

Turning now to the second component $\bar{x}_2$, we see by exactly the same type of argument that it is precisely the second component of the exact solution of a neighboring system.

$$(U+\delta U^{(2)})x^{(2)} = c + \delta c^{(2)}. \qquad (4.9)$$

The matrix $\delta U^{(2)}$ is null in its first row. (The first equation is not involved in the reduction of the second equation.) Similarly the first component of $\delta c^{(2)}$ is zero. We certainly have

$$|\,\delta U^{(2)}\,| \le (n-1)\epsilon\,|\,U\,|,\ |\,\delta c^{(2)}\,| \le (n-1)\epsilon\,|\,c\,| \qquad (4.10)$$

and hence a fortiori

$$|\,\delta U^{(2)}\,| \le n\epsilon\,|\,U\,|,\ |\,\delta c^{(2)}\,| \le n\epsilon\,|\,c\,|. \qquad (4.11)$$

In general $\bar{x}_r$ is precisely the rth component of the exact solution $x^{(r)}$ of a neighboring system

$$(U+\delta U^{(r)})x^{(r)} = c + \delta c^{(r)}, \qquad (4.12)$$

where certainly $\delta U^{(r)}$, and $\delta c^{(r)}$, which are null in their first $r-1$ rows, satisfy

$$|\,\delta U^{(r)}\,| \le n\epsilon\,|\,U\,|,\ |\,\delta c^{(r)}\,| \le n\epsilon\,|\,c\,|. \qquad (4.13)$$

The essential difference between solving the triangular system $Ux = c$ by Gauss-Jordan and by back-substitution is that whereas for the latter the whole of the computed solution is the exact solution of a single neighboring system

$$(U+\delta U)x = c + \delta c \qquad (4.14)$$

(indeed it is easy to avoid having any perturbation δc), with the former each *component* belongs to the exact solution of a neighboring system *but it is a different neighboring system for each one.* We now analyze the consequences of this last remark. If x is the exact solution of $Ux = c$, then if

$$(U+\delta U^{(r)})x^{(r)} = c + \delta c^{(r)} \qquad (4.15)$$

we have

$$\begin{aligned}
x^{(r)} &= (U+\delta U^{(r)})^{-1}c + (U+\delta U^{(r)})^{-1}\delta c^{(r)},\\
&= x + e^{(r)} + f^{(r)}\ \text{(say)},
\end{aligned} \qquad (4.16)$$

where

$$\|\,e^{(r)}\,\| \le \left\{\frac{\|\,U^{-1}\,\|\,\|\,\delta U^{(r)}\,\|}{1 - \|\,U^{-1}\,\|\,\|\,\delta U^{(r)}\,\|}\right\}\|\,x\,\|, \qquad (4.17)$$

$$\|f^{(r)}\,\| \le \frac{\|\,U^{-1}\,\|\,\|\,\delta c^{(r)}\,\|}{1 - \|\,U^{-1}\,\|\,\|\,\delta U^{(r)}\,\|}. \qquad (4.18)$$

Remembering that $c = Ux$ and therefore

$$\|\,c\,\| \le \|\,U\,\|\,\|\,x\,\|,$$

we find

$$\begin{aligned}
&\|\,e^{(r)} + f^{(r)}\,\|\\
&\qquad \le \frac{n\epsilon[\|\,U^{-1}\,\|\,\|\,U\,\| + \|\,U^{-1}\,\|\,\|\,U\,\|]\,\|\,x\,\|}{1 - n\epsilon\,\|\,U^{-1}\,\|\,\|\,U\,\|}.
\end{aligned} \qquad (4.19)$$

If we use the l_∞ norms we have

$$\begin{aligned}
\frac{\|\,x^{(r)} - x\,\|_\infty}{\|\,x\,\|_\infty} &\le 2n\epsilon\left(\frac{\|\,U^{-1}\,\|_\infty\,\|\,U\,\|_\infty}{1 - n\epsilon\,\|\,U\,\|_\infty\,\|\,U^{-1}\,\|_\infty}\right)\\
&= 2n\epsilon\,\frac{\kappa}{1 - n\epsilon\kappa}.
\end{aligned} \qquad (4.20)$$

In all cases these inequalities hold only under the assumption that $n\epsilon\kappa < 1$. Now since $\bar{x}_r = x_r^{(r)}$, we certainly have

$$\begin{aligned}
|\,\bar{x}_r - x_r\,| = |\,x_r^{(r)} - x_r\,| &\le \max_i |\,x_i^{(r)} - x_i\,|\\
&= \|\,x^{(r)} - x\,\|_\infty
\end{aligned} \qquad (4.21)$$

and hence

$$\begin{aligned}
\|\,\bar{x} - x\,\|_\infty &= \max |\,\bar{x}_r - x_r\,|\\
&\le \max \|\,x^{(r)} - x\,\|_\infty \le \frac{2n\epsilon\kappa}{1 - n\epsilon\kappa}\|\,x\,\|_\infty.
\end{aligned} \qquad (4.22)$$

This result is precisely what we would have obtained had the same δU and δc given all the computed components.

Now when back-substitution is used to solve a triangular set, it is well known (see e.g. [3], pp. 99–107) that the computed solution is often more accurate than one would expect, having regard to the size of the δU and δc derived by a backward-error analysis. However, this is not of great importance here. Remember that we are primarily interested in the solution of a system of equations with a full, square matrix, and the solution of the triangular system is merely the second half of the process. In going from the original system to the triangular system, errors comparable with these corresponding to (4.22) above will already have been made.

Table I.

Original System				
	MATRIX			R.H.S.
.826354	.432175	.613256	.614227	.722872
	.000547	.814712	.816328	.154248
		.915316	.814275	.109844
			.982176	.602286
Equation 1 After First Reduction				
.826354	.000000	−643.076	−644.352	−121.146
Equations 1 and 2 After Second Reduction				
.826354	.000000	.000000	−72.2644	−43.9726
	.000547	.000000	.0915516	.0564772
Final Diagonal System				
.826354	.000000	.000000	.000000	.341074
	.000547	.000000	.000000	.000336315
		.915316	.000000	−.389482
			.982176	.602286

Table II.

Solutions, Errors and Residuals

Gauss-Jordan		Back-substitution		Solution correct to 6 figures
Solution	Error	Solution	Error	
0.412746	−0.000409	0.413503	0.000348	0.413155
0.614835	−0.000092	0.614260	−.000667	0.614927
−0.425516	0.000001	−0.425516	.000001	−0.425517
0.613216	0.000000	0.613216	.000000	0.613216
Residual		Residual		
0.000378··		0.000000742··		
−0.000000714··		−0.000000399··		
−0.000000855··		−0.000000855··		
−0.000000038··		−0.000000038··		

Hence the rather exceptional accuracy often obtained in back-substitution avails us little.

We may summarize this by saying that when we solve a square system $Ax = b$ by Gaussian elimination the computed solution is the exact solution of some neighboring system $(A+E)x = b$, and the bound for E does not involve κ. When it is solved by Gauss-Jordan, the computed solution is *not* the exact solution of such a neighboring system but the error $\| x_c - x \|$ is of just the same order of magnitude as that corresponding to an x_c, which *is* the solution of such a system.

An analogous situation has already been diagnosed in the case of matrix inversion by Gaussian elimination and back-substitution. It is *not* the case that the computed inverse X is the exact inverse of some $(A+E)$ where $\| E \|$ has a bound which does not involve κ. It is true, however, that the rth column x_r is the rth column of the exact inverse of some neighboring $(A+E_r)$, but it is a different E_r for each column.

Turning now to the residual, the fact that it is a different $\delta U^{(r)}$ for each component is quite serious in its implications, and the residuals corresponding to the Gauss-Jordan solution are often larger than those corresponding to back-substitution by a factor of order κ. Note that this merely means that the Gauss-Jordan solution gives a residual which is commonly of the order of magnitude one naturally associates with its accuracy; the solution by back-substitution gives a much *smaller* residual then one would expect, and this performance is attained only because of a special correlation in the errors.

5. Numerical Example

The points made above are illustrated by a simple example of order four having just one small pivot. In Table I we give the successive steps in the Gauss-Jordan reduction. The computation was done in 6-digit floating-point decimal arithmetic, but for easier recognition standard floating-point notation is not used. Equations which are unmodified in any reduction stage are not repeated. Observe that in the first stage of reduction, growth by a factor of 1,000 occurs in elements of the first equation as a result of the use of the pivot $u_{22} =$ .000547. In Table II we give the computed solutions obtained with Gauss-Jordan and Gaussian elimination respectively, and for comparison we give also the correctly rounded solution. The errors are of the order of magnitude to be expected having regard to the condition number of the triangular matrix; the back-substitution gave marginally larger errors. (Note that for fairer comparison, back-substitution was done without accumulation of inner-products.) Turning now to the residuals, we see that the first residual corresponding to Gauss-Jordan is far larger than that corresponding to back-substitution. The large component of the residual arises in the first equation, and the backward error analysis of Section 3 forecasts this since for this example it is in the first equation that we have the large components of E.

Acknowledgments. We would like to thank Professor G.H. Golub for drawing our attention to this problem and for stimulating discussions on the topic.

References

1. Forsythe, G.E., and Moler, C.B. *Computer Solution of Linear Algebraic Systems.* Prentice-Hall, Englewood Cliffs, N.J., 1967.
2. Wilkinson, J.H. Error analysis of direct methods of matrix inversion. *J. ACM 8*, 3 (July 1961), 281–330.
3. Wilkinson, J.H. *Rounding Errors in Algebraic Processes.* Her Majesty's Stationery Store, London; and Prentice-Hall, Englewood Cliffs, N.J., 1963.
4. Wilkinson, J.H. *The Algebraic Eigenvalue Problem.* Oxford University Press, London, 1965.